Information theory, evolutionary innovations and evolvability

Wagner, Andreas

Abstract: How difficult is it to 'discover' an evolutionary adaptation or innovation? I here suggest that information theory, in combination with high-throughput DNA sequencing, can help answer this question by quantifying a new phenotype's information content. I apply this framework to compute the phenotypic information associated with novel gene regulation and with the ability to use novel carbon sources. The framework can also help quantify how DNA duplications affect evolvability, estimate the complexity of phenotypes and clarify the meaning of 'progress' in Darwinian evolution. This article is part of the themed issue 'Process and pattern in innovations from cells to societies'.

DOI: https://doi.org/10.1098/rstb.2016.0416

Information theory, evolutionary innovations, and evolvability

Andreas Wagner[a,b,c]

[a]University of Zurich, Institute of Evolutionary Biology and Environmental Studies, Zurich, Switzerland,

[b]Swiss Institute of Bioinformatics, Lausanne, Switzerland,

[c]Santa Fe Institute, Santa Fe, New Mexico, USA

Abstract

How difficult is it to "discover" an evolutionary adaptation or innovation? I here suggest that information theory, in combination with high throughput DNA sequencing, can help answer this question by quantifying a new phenotype's information content. I apply this framework to compute the phenotypic information associated with novel gene regulation, and with the ability to utilize novel carbon sources. The framework can also help quantify how DNA duplications affect evolvability, estimate the complexity of phenotypes, and clarify the meaning of "progress" in Darwinian evolution.

Keywords: evolvability, gene duplication, progress

Introduction

Evolutionary biologists have a long-standing interest in information theory, because it is ultimately information encoded in DNA that renders the survivors of natural selection well-adapted to their environment [1-4]. Among the first researchers to explore the link between information and evolution was Motoo Kimura. He built on earlier work by J.B.S. Haldane to argue that adaptive evolution accumulates genetic information in proportion to the rate at which alleles are replaced by better-adapted alleles [5, 6].

More recently, two independent lines of research have connected evolutionary biology and information theory. The first is centered on organisms and their *phenotypes*, which may harbor information about the environment [3, 7-14]. For example, the growth rate of bacteria depends on information that cells sense about environmental nutrients [3, 7, 8, 10, 12, 14, 15]. The second line focuses on *genotypes* [2, 16-19], where information-theoretic concepts such as Shannon's entropy [2, 18] can help recast equations from classical population and quantitative genetics, to describe changes of genotypes, allele frequencies, and fitness in information-theoretic terms. This line of research shows that natural selection can increase information encoded in the distribution of a population's allele frequencies [2, 17].

Experimental evolution, be it through *in vitro* selection [20-23], through directed evolution of macromolecules [24-28], or through laboratory evolution of organisms [29], is a powerful tool to discover novel phenotypes, such as the ability to resist novel antibiotics, to regulate old genes in new ways, and to thrive on novel sources of energy. A fundamental question about evolutionary adaptations and innovations – qualitatively new and beneficial phenotypes – is how difficult it is to acquire or "discover" them. For example, it may be easier to acquire the ability to extract energy from some novel nutrient, if this ability requires only one new enzyme (biochemical reaction) instead of two or more. I here suggest that basic concepts from information theory, together with data from high-throughput DNA sequencing technologies may help us answer this and related questions quantitatively for different kinds of phenotypes.

In the next section, I will first introduce a suitable information-theoretic framework, and illustrate its use to understand evolution by DNA or gene duplication. Second, I will apply the framework to two different kinds of phenotypes, the DNA binding phenotypes of transcriptional regulators [30, 31], and the metabolic phenotypes that allow an organism to procure energy and manufacture essential biomass molecules [32]. Third and finally, I will show how sequence data from experimental evolution could help quantify differences in the amount of information gained by different evolutionary adaptations.

Results

All evolutionary adaptations and innovations originate in some space of genotypes. Evolving populations of organisms or molecules explore genotype spaces through DNA mutation, genetic drift, and selection. The relationship between genotypes and phenotypes – the genotype-phenotype map – has been studied for multiple different kinds of genotype spaces, either exhaustively (for small spaces) or through random sampling, using both computational and experimental techniques [20, 21, 33-38]. Such efforts show that, first, astronomically many genotypes usually form the same phenotype, and these genotypes are organized into one or more networks connected by point mutations. Second, the genotype networks of different phenotypes are interwoven in complex ways [35, 38, 39]. Third, some phenotypes have larger genotype networks than others. This observation is important to understand phenotypic evolvability, the ability of an organism with a specific phenotype to bring forth novel phenotypes through DNA mutations [40]. Both computational analyses and empirical data show that populations evolving on large genotype networks are – with possible exceptions [41, 42] – more likely to "discover" new and beneficial phenotypes, because such populations can explore a larger proportion of genotype space [38, 40, 43, 44].

For any observed phenotype *P*, such as a protein's ability to bind or react with a specific molecule, I denote the set of genotypes with this phenotype as G_P. For simplicity, I focus on discrete, qualitative phenotypes (e.g., binding or not) rather than on quantitative phenotypes (e.g., binding with a specific affinity), thus assuming that all genotypes with a particular phenotype are equivalent. Consider first a genotype space G, and the Shannon entropy of a random variable that assumes values $g \in G$ with equal probability $\frac{1}{|G|}$, where $|G|$ denotes the number of genotypes in G. This entropy computes as $H(G) = -\sum_{g \in G} \frac{1}{|G|}\left(log_2 \frac{1}{|G|}\right) = -log_2 \frac{1}{|G|} = log_2|G|$ [45]. The Shannon entropy for the same random variable defined on a subset G_P of genotypes with a specific phenotype *P* (Figure 1a), computes analogously as

$$H(G_P) = -\sum_{g \in G_P} \frac{1}{|G_P|}\left(log_2 \frac{1}{|G_P|}\right) = -log_2 \frac{1}{|G_P|} = log_2|G_P|$$

These observations give rise to the following definition.

Definition 1: The information content of phenotype *P* is given by

$$I(P) := log_2|G| - log_2|G_P| \qquad (1)$$

Equivalently, if $|G_P| = f_P|G|$, where f_P indicates the fraction of genotype space occupied by G_P, then $I(P) = -log_2\ f_P \geq 0$. Analogous quantities have been called self-information, functional information, surprisal, and (biological) complexity in other contexts [45-48]. The greater a phenotype's information content is, the more information is required to encode this phenotype. To compare data from genotype spaces of different dimensions (e.g., proteins of different length L), it can be useful to consider information content per monomer ($I(P)/L$).

Some empirical data on phenotypic information content is available for macromolecules. For example, in vitro selection experiments identifying ATP-binding proteins from a random library of proteins with 80 amino acids show that a fraction f_P=10^{-11} or $|G_P|$ =$20^{80}\times10^{-11}\approx1.2\times10^{93}$ proteins of this length can bind ATP [21]. The amount of information associated with the ATP-binding phenotype is $I(P) = log_2|20^{80}|$-$log_2|20^{80} \times 10^{-11}|$=$-log_2|10^{-11}|$=36.6 bits, which is much lower than the amount of information needed to specify a single amino acid sequence ($log_2(20^{80}) \approx 345.8$ bits), because many proteins can bind ATP.

Unlike in vitro selection, laboratory evolution experiments often do not start from random collections of genotypes, but from genotypes that already have a specific phenotype P_{Old} and acquire a novel phenotype P_{New} (Figure 1b). For example, in a directed evolution experiment, TEM-1 β-lactamase molecules that convey resistance to ampicillin may acquire the ability to cleave the antibiotic cefotaxime. Denote as G_{Old} the subset of genotypes with the old phenotype, and as G_{New} the subset of genotypes with the new phenotype.

Definition 2: The information *change* associated with the acquisition of a new phenotype P_{New} starting from some phenotype P_{Old}, is given by

$$\Delta I \coloneqq I(P_{New}) - I(P_{Old}) = log_2|G_{Old}| - log_2|G_{New}| \quad (2)$$

Here, one can distinguish two different scenarios. In the first, individuals with the new phenotype have also preserved the old phenotype (Figure 1b), which implies that $G_{New} \subseteq G_{Old}$, and $\Delta I \geq 0$. In this case, ΔI is equivalent to a Kullback-Leibler distance or relative entropy, an important quantity in information theory [45]. (Supplementary Results 1). In the second scenario, G_{New} is not a proper subset of G_{Old} (Figure 1c). For example, consider β-lactamase enzymes that have evolved the ability to inactivate cefotaxime, but that may not have retained the old ability to inactivate ampicillin. In this case ΔI can be negative, for example if more genotypes encode the ability to cleave cefotaxime than ampicillin.

To illustrate one potential use of this framework, consider DNA duplication, which has long been thought to increase evolvability [49-51]. To help quantify the advantage of duplicated DNA over single-copy DNA in exploring a genotype space, consider some phenotype, such as a

regulatory region's ability to bind a transcription factor, or a protein's ability to catalyze a specific chemical reaction, and the set of genotypes G_P associated with this phenotype. When the DNA encoding this phenotype becomes duplicated, both copies can undergo DNA mutation independently. Thus, they evolve in a larger genotype space, which comprises many more ($|G|^2$) genotypes.

Such duplication can affect the information content of *P,* if only one of two copies of the duplicated DNA is sufficient to encode *P*. In this case, the difference between phenotypic information content after and before duplication is equal to

$$\frac{1}{2L} log_2 \frac{f_P}{2-f_P} \approx \frac{1}{2L} (log_2 f_P - 1) < 0 \qquad (3)$$

per nucleotide or any other suitable unit, such as an amino acid monomer (See Supplementary Results 2, including for an analogous calculation of changes in absolute information content). Here *L* is the dimension of *G* (e.g., the pre-duplication length of DNA). The approximation holds for $f_P \ll 1$. Importantly, this quantity is negative: Duplication decreases a phenotype's information content per nucleotide, because the set of post-duplication genotypes with phenotype *P* occupy a larger fraction of genotype space. This is important, because such a larger fraction of genotypes is associated with higher evolvability [38, 40, 43, 44]. Thus, information theory can help link DNA duplication and evolvability. The set of genotypes associated with *P* expands after duplication by a factor $(2 - f_P)/f_P \approx 2/f_P$ (Supplementary Results 2). Because this expansion factor scales as $1/f_P$, duplication will enhance evolvability to the greatest extent for phenotypes formed by few genotypes (small f_P). In terms of the ATP-binding protein example above, where $|G|$ =20^{80}, f_P=10^{-11}, and *L*=80, duplication would reduce the phenotypic information content by $log_2 f_P - 1 = (log_2 10^{-11} - 1)/160 \approx -0.23$ bits per amino acid.

Transcription factor binding phenotypes. Numerous evolutionary adaptations and innovations have been associated with the origin of new gene regulation mediated by new transcription factor binding sites on DNA, from changes in pathogen virulence to new body plans, such as the origin of two-winged insects [52-54]. I next analyze a genotype space of 4^8 = 65,536 DNA sequences of length eight nucleotides to illustrate the information change associated with new transcription factor binding sites. To this end, I take advantage of previously published protein binding microarray experiments that measured how strongly each of 187 mouse transcriptional regulators binds to all sequences in this space [30, 31] (Supplementary Methods).

The phenotypes I analyze here are a DNA sequence's ability to bind specific regulators. For a *de novo* origin of transcription factor binding, the relevant phenotypic information content is that of a binding site (definition 1). Among the 187 regulators, this content varies widely (Figure 2a;

$I(P)$ =4.48- 9.31 bits, median: 5.72 bits), because the fractional volume of genotype space bound varies among regulators. Binding sites with lower information content would be easier to acquire *de novo* [55, 56]. The frequently made assumption that individual nucleotides contribute additively to phenotypic information [48, 57, 58] can lead to substantial underestimation of phenotypic information, i.e., by up to 8.22 bits (300-fold in terms of f_P, Figure S1, Supplementary Results 3, Supplementary Methods).

If a site gets duplicated, such that the two duplicates evolve separately, and only one of them needs to preserve regulator binding, the change in phenotypic information content as a result of duplication lies between -0.34 and -0.64 bits per nucleotide (equation 3, with data from Figure 2a, where $0.0016<f_P<0.045$), meaning that 44.4-1250 times more genotypes can be explored. In reality, duplication will confer an even greater advantage, because often entire regulatory regions and not just individual binding sites are duplicated.

When a binding site for a new regulator originates from one for an old regulator, *and* if binding of the old regulator is preserved (for example, because the old regulator directs essential gene expression in a different tissue), then phenotypic information increases (definition 2, Figure 2b). For the 187 regulators considered here, the minimal increase is 0.04 bits when a binding site for factor Myb originates from one for factor Mybl1, because these regulators belong to the same gene family, and 97.2% of the 1969 sites bound by Mybl1 are also bound by Myb. The largest increase (11.5 bits) occurs when binding sites for transcription factor Mnt emerge from those for Sp110, because Sp110 binds to 2933 sites, but only one of them is also bound by Mnt. The complexity increase is generally lower if the old site had high information content (Spearman's $r=-0.22$; $P<10^{-17}$; $n=29290$, Figure 2b, inset), which shows that phenotypic information changes can depend on ancestral phenotypes and are thus contingent on evolutionary history.

If binding to an old transcriptional regulator need not be preserved after a new binding phenotype originates, then the distribution of information change is symmetric (Figure 2c), because for every value of information change X that occurs when binding is gained by some new regulator Y and lost by an old regulator Z, there is an opposite value –X when binding by Z is gained and binding by Y is lost. The maximal information loss or gain is 4.83 bits (Sp110-binding originating from Usf1-binding). Its minimum is zero for regulator pairs (e.g., Hbp1 and Rfx4) that bind the same number of sites.

Metabolic genotypes and phenotypes. The metabolic genotype of an organism comprises all genes encoding metabolic enzymes. Systems biologists often represent this genotype more compactly, on the level of metabolic *reactions* these enzymes catalyze, by the presence or absence of specific reactions from a known "universe" of such reactions [59, 60] (Figure S3a).

Such a genotype encodes a biochemical reaction network that transforms environmental nutrients into essential biomass molecules, such as amino acids and nucleotides. A metabolic genotype is *viable* only if it can produce all biomass molecules an organism needs in a given nutrient environment. One can compute viability for any known genotype under some simplifying assumptions [59], and these predictions are often in good agreement with experimental observations [60-63].

The metabolic genotypes of free-living organism like *E.coli* are members of a vast genotype space that can only be explored by sampling. I restrict myself here to a much smaller universe of 45 metabolic reactions from central carbon metabolism, which gives rise to a more tractable genotype space [64] (Figure S2, Supplementary Methods). Given the right nutrients, the biochemical network encoded by these reactions can manufacture 13 essential biomass precursors, such as ribose 5-phosphate and oxaloacetate (Figure S2). The metabolic genotype space I explore is formed by all possible (2^{45}) subsets of these reactions. I consider 10 minimal chemical environments that differ only in the sole carbon source they contain (Supplementary Methods, Figure S3b), and represent a metabolic phenotype as the combination of carbon sources on which a metabolism is viable (Figure S3b and 3c). Considering all possible combinations of 10 carbon sources on which a metabolism could be viable, this leads to 2^{10} possible metabolic phenotypes.

In an earlier contribution, we have exhaustively computed these metabolic phenotypes for all $2^{45}\approx10^{13}$ metabolic genotypes [64], which allows me to analyze their phenotypic information content. Some phenotypes contain much less information than others, e.g., viability on fructose and glucose require 14.8 bits of information, whereas viability on all 10 carbon sources except glutamate and α-ketoglutarate requires 28.6 bits. Starting from a metabolic phenotype, viability on an additional carbon source requires an average of 0.75 additional bits (Figure S3d, inset). Neglecting non-additive interactions among reactions underestimates phenotypic information (Figure S3e, S4d), and duplication causes a substantial reduction in information by up to 29 bits (Supplementary Results 4). The distribution of information gain and information change are broad (Figure S3f, S3g, Supplementary Results 4-6). Gaining viability on a given carbon sources can be informationally cheap (e.g., α-ketoglutarate) or expensive (acetate). Perhaps surprisingly, gaining viability on some carbon sources may lead to reduced phenotypic information, as a result of complex correlations between phenotypes (Supplementary Results 4).

Inferring information content from sequence data. Tractable genotype spaces like those I discussed so far are the exception. Usually, astronomically many genotypes encode the same phenotype, and because it is impossible to identify all of them, one cannot infer the information content of any one phenotype (definition 1). What is more, sequencing technology does not

simply enumerate genotypes but *samples* them from an evolving population. I will argue next that it may nonetheless be possible to estimate the information *change* associated with a novel phenotype (definition 2). In doing so, I make simplifying assumptions whose relaxation will require future work. My main point is that quantifying phenotypic information change may be within reach of current technologies.

Consider two populations of which one is well-adapted to some ancestral environment (with phenotype P_{Old}), and another one is adapted to a new environment, such as one that harbors a novel nutrient, an antibiotic, or another stressor, and thus requires an altered phenotype P_{New}. I assume that both populations comprise asexually reproducing haploid individuals, that they are in mutation-selection-drift balance subject to Wright-Fisher dynamics [65], and that they have equal effective sizes N_e and mutation rates μ (per genome and generation). I also assume that both phenotypes P_{Old} and P_{New} are subject to strong truncation selection, that is, mutations that disrupt each phenotype are lethal. The two phenotypes may differ in their numbers of associated genotypes G_{Old} and G_{New}, and thus also in their information content. The task is to estimate this difference ($log_2|G_{Old}| - log_2|G_{New}|$) from two samples of n genotypes (DNA sequences), one from each of the populations. I model this difference as a difference in the average rate of strongly deleterious (lethal) mutations across all genotypes, or equivalently, in the average rate of neutral mutations. If l_{Old} and l_{New} denote the average proportion of all strongly deleterious (lethal) mutations in the two populations, then the average neutral mutation rate becomes $\mu_{Old} = (1 - l_{Old})\mu$ and $\mu_{New} = (1 - l_{New})\mu$. Assuming further that mutations in all viable genotypes are equally likely to be strongly deleterious, one obtains the relationships $|G_{Old}| = \mu_{Old}|G|$ and $|G_{New}| = \mu_{New}|G|$, where $|G|$ is the total size of genotype space. It is then easy to see that

$$log_2|G_{Old}| - log_2|G_{New}| = log_2(\frac{\mu_{Old}}{\mu_{New}}) = log_2|2N_e\mu_{Old}| - log_2|2N_e\mu_{New}| \quad (5)$$

Thus, estimating the difference in phenotypic information content requires estimating the quantities $\theta_i = 2N_e\mu_i$, which are of broad importance in population genetics because they predict a population's amount of neutral polymorphisms [65, 66]. If P_{New} harbors more information than P_{Old} ($|G_{Old}| > |G_{New}|$), then $2N_e\mu_{Old} > 2N_e\mu_{New}$, and the population with P_{New} would harbor more alleles.

A maximum likelihood estimator of θ_i is the number of *different* genotypes k_i in a random sample of n genotypes sequenced from the populations [67]. Importantly, the sampling distribution of k_i is known, and it can help infer the minimal difference in information content detectable from a sample of n sequences (Supplementary Methods). Specifically, one can compute the probability of falsely rejecting the null-hypothesis that phenotype P_{New} harbors

more information than P_{Old}. Figure 3 shows the minimally detectable information difference (see legend), for multiple values of n and $\theta_{New} = 2N_e\mu_{New}$. White regions in the plot indicate that the information content of two phenotypes is indistinguishable. In a region of the plot where the test can discriminate at least x bits, p<0.05 for all values of θ_{Old}, such that $\theta_{Old} > 2^x\theta_{New}$.

In this analysis, I did not explore populations with $\theta_{Old}, \theta_{New} < 1$, because such populations are monomorphic most of the time [66], which implies that even when sequencing multiple genotypes, most of the genotypes would be genetically identical. At the other extreme are values of $\theta_{New} \gg n$ (and thus also $\theta_{Old} \gg n$), where one cannot discriminate the information content of two phenotypes (Figure 3), because both populations are so highly polymorphic that all n sampled sequences may be different from each other. Thus, best discrimination between the information content of two phenotypes requires that $1 \ll \theta_{Old}, \theta_{New} \ll n$ (lower right corner of Figure 3).

Discussion

The genotypes encoding a new phenotype may be difficult to access by an evolving population for two reasons. First, the phenotype may have high information content, implying that the few genotypes encoding it may be difficult to find through random search in a vast genotype space. Second, the population may be distant from these genotypes, requiring multiple genetic changes or inviable mutational intermediates to reach them. The information-theoretic framework eliminates the latter, historical factors from consideration, which is both an advantage and a limitation. In the data I analyzed here, most phenotypes can be reached through few genetic changes (Supplementary Methods), but this may not hold in larger genotype spaces [68].

I have restricted myself here to qualitative or threshold phenotypes (binding/non-binding, viability/non-viability). They have proven useful in past experimental estimates of phenotypic information, such as that of RNA ligase ribozymes (L=220) whose $I(P)$ can be estimated at 43.2 bits [20]. Limited empirical data is also available about the information content of quantitative phenotypes. For example, a 10-fold increase in an RNA aptamer's binding affinity to guanosine triphosphate (GTP) requires 10 additional bits of information [47, 58]. However, extending the concepts of this paper to quantitative phenotypes remains a task for future work (Supplementary Results S1).

Qualitative phenotypes are simplifications, but they also help separate properties intrinsic to a phenotype from properties of a population harboring a phenotype [48, 69]. The latter depends not only on phenotypic information, but also on many factors affecting a population's dynamics, such as (effective) population size N_e and genomic mutation rate μ. For example, if $N_e\mu \ll 1$,

then all members of a population have identical genotypes most of the time [66]. A phenotype's information content estimated from such a monomorphic population would be $log_2|G|$, a highly misleading value, because the population does not harbor any of the myriad other genotypes that might encode the phenotype. Likewise, during adaptive evolution, apparent phenotypic information can rise dramatically and transiently before reaching a mutation-selection equilibrium, for example while an adaptive mutant genotype goes to fixation [48, 69].

A typical experiment to estimate phenotypic information changes from an evolving population would start at the end-point of a previous (laboratory) evolution experiment, in which a population has adapted evolutionarily to a novel environment, such as one containing a novel nutrient or antibiotic. The experiment would then establish two evolving populations, one derived from a single pre-evolution genotype and evolved in the ancestral environment (e.g., without antibiotic), the second from a single post-evolution (adapted) genotype, and evolved in the novel environment (e.g., with antibiotic). After each population has evolved sufficiently long to reach approximate mutation-selection balance, one would sequence n randomly chosen individuals from each population, and infer ΔI from the number of different alleles (genomes) sampled. As I argued, to best discriminate between phenotypic information content in both populations, one needs $n \gg N_e\mu \gg 1$. This is entirely feasible, even with multi-megabase genomes. For example, in *E.coli*, where $\mu \approx 10^{-3}$ [70], a population of $N_e=10^4$ individuals yields $N_e\mu \approx 10$. Current technology permits sequencing of more than 100 clones isolated from such a population, such that $n > 100 \gg N_e\mu \gg 1$.

My argument that today's sequencing technology can help distinguish even modest phenotypic information changes rests on simplifying assumptions, among them that a sampled population should not be far from mutation-selection balance. Such a balance is approached exponentially with decay parameter $\lambda = (1 + 4N_e\mu)/2N_e$ [66, p 204]. For an evolving *E.coli* population of 10^4-10^7 individuals with $\mu \approx 10^{-3}$ mutations per genome and generation [70], the half-life of this decay is given by $\ln 2 / \lambda = \ln 2\,(2N_e)/(1 + 4N_e\mu) \approx$ 340-370 generations, well within the time scale of a laboratory evolution experiment. Other assumptions, such as that of truncation selection, unbiased mutational sampling of G_P, as well as a uniform deleterious mutation rate for all genotypes in G_P, will need to be relaxed in more sophisticated modeling work, which will also be required for a rigorous sampling theory estimating quantities such as confidence intervals for phenotypic information changes.

The information-theoretic framework can speak to broad and fundamental questions in evolutionary biology. One of them is whether some organisms and phenotypes are more evolvable than others. Here, information theory unifies previous observations [38, 40, 43, 44] to

show that phenotypes with low information content are more evolvable (with possible exceptions, where genotypes form highly fragmented sets in genotype space, Figure 1d [44, 71]). Relatedly, information theory can also help quantify the extent to which DNA or gene duplications increase evolvability (equation 3). In addition, the framework can help solve the recalcitrant problem of how to define the complexity of phenotypes and organisms: More complex phenotypes are those with higher phenotypic information content. Relatedly, it can help answer under what circumstances evolution implies "progress". This is controversial, partly because adaptive evolution can be regressive and lead to trait loss [72]. With a definition of phenotypic information in hand, evolutionary progress can be defined as an increase in phenotype information content in an evolving lineage.

Acknowledgments

I would like to thank Joshua Payne, José Aguilar Rodrigues, and Rzgar Hosseini for valuable discussions on the data analyzed here, as well as Troy Day and an anonymous reviewer for valuable comments. I also acknowledge support by Swiss National Science Foundation grant 31003A_146137, by an EpiphysX RTD grant from SystemsX.ch, as well as by the University Priority Research Program in Evolutionary Biology at the University of Zurich.

Figure Captions

Figure 1. Sets of genotypes with the same phenotype can have various topological relationships. Large rectangles symbolize genotype space, circles correspond to genotypes, and straight lines connect 1-mutant neighbors, i.e., genotypes that differ by a small genetic change such as a single nucleotide change. Each set of genotypes is shown as a network, because such sets form networks in genotype space. **a)** A hypothetical set (network) of genotypes with the same phenotype. The set is shown as a single genotype network, but I note that it could consist of multiple disconnected networks. **b)** Two sets of genotypes, the first associated with an old phenotype (black *and* grey circles), the second with a new phenotype (grey circles only, a subset of the first set). **c)** Sets of genotypes with an old phenotype (black circles), a new phenotype (white circles), or with both an old and a new phenotype (grey circles). Unlike in b), the genotype set of the new phenotype is not a subset of the genotype set with the old phenotype. **d)** The sets of genotypes encoding different phenotypes can be non-overlapping.

Figure 2. Phenotypic information associated with new transcription factor binding. Data are based on experimentally measured binding of 187 mouse transcription factors to all possible DNA binding sites of length eight [30, 31, 73]. **a)** Histogram of the information content of the DNA binding phenotype of each transcription factor (definition 1 and equation 1). **b)** The gain in information content associated with acquisition of a new DNA binding phenotype, when an old phenotype is simultaneously preserved (equation 2). The inset shows this gain in information content (vertical axis) as a function of the information content of the old phenotype (horizontal axis). Circles correspond to means, boxes to standard errors, and whiskers indicate 95 percent confidence intervals. Data in b) are based on all 29290 pairs of transcription factors whose sets of binding sites overlap. **c)** The change in information content associated with acquisition of a new DNA binding phenotype when the old phenotype need not be simultaneously preserved (equation 4). The red line indicates the fit to a Gaussian distribution. Data in c) are based on all 187^2 pairs of transcription factors in the data set.

Figure 3. High-throughput sequencing can help distinguish even modest differences in phenotypic information content. Minimally distinguishable information content of two phenotypes (in bits), color-coded as indicated in the legend, for a given sequence coverage n (horizontal axis), and a given value of $\theta_{New} = 2N_e\mu_{New}$. To create this plot, I chose multiple values of n and θ_{New}, and determined the minimal value of θ_{Old}, θ_{Old}^{min} such that $p = \sum_{-(n-1)\leq\Delta k\leq 0} Prob\left(\Delta K = \Delta k \middle| n, \theta_{Old}^{min}, \theta_{New}\right) < 0.05$ (see Methods) for each of these values. The minimally detectable information difference is then given by $log_2\left|\theta_{Old}^{min}\right| - log_2|\theta_{New}|$.

References

[1] Maynard-Smith, J. 2000 The concept of information in biology. *Philosophy of Science* **67**, 177-194. (DOI:10.1086/392768).
[2] Frank, S. A. 2012 Natural selection. V. How to read the fundamental equations of evolutionary change in terms of information theory. *Journal of Evolutionary Biology* **25**, 2377-2396. (DOI:10.1111/jeb.12010).
[3] van Baalen, M. 2013 Biological information: why we need a good measure and the challenges ahead. *Interface Focus* **3**. (DOI:10.1098/rsfs.2013.0030).
[4] Wallace, R. & Wallace, R. G. 1998 Information theory, scaling laws and the thermodynamics of evolution. *Journal of Theoretical Biology* **192**, 545-559. (DOI:10.1006/jtbi.1998.0680).
[5] Kimura, M. 1961 Natural selection as a process of accumulating genetic information in adaptive evolution. *Genetical Research* **2**, 127-&.
[6] Haldane, J. B. S. 1957 The cost of natural selection. *Genetics* **55**, 511-524.
[7] Hoffmann, R. J. 1978 Environmental uncertainty and evolution of physiological adaptation in Colias butterflies. *American Naturalist* **112**, 999-1015. (DOI:10.1086/283343).
[8] Donaldson-Matasci, M. C., Bergstrom, C. T. & Lachmann, M. 2010 The fitness value of information. *Oikos* **119**, 219-230. (DOI:10.1111/j.1600-0706.2009.17781.x).
[9] Donaldson-Matasci, M. C., Bergstrom, C. T. & Lachmann, M. 2013 When Unreliable Cues Are Good Enough. *American Naturalist* **182**, 313-327. (DOI:10.1086/671161).
[10] McNamara, J. M. & Dall, S. R. X. 2010 Information is a fitness enhancing resource. *Oikos* **119**, 231-236. (DOI:10.1111/j.1600-0706.2009.17509.x).
[11] Lachmann, M. & Bergstrom, C. T. 2004 The disadvantage of combinatorial communication. *Proceedings of the Royal Society B-Biological Sciences* **271**, 2337-2343. (DOI:10.1098/rspb.2004.2844).
[12] Rivoire, O. & Leibler, S. 2011 The value of information for populations in varying environments. *Journal of Statistical Physics* **142**, 1124-1166. (DOI:10.1007/s10955-011-0166-2).
[13] Tkacik, G. & Bialek, W. 2016 Information processing in living systems. In *Annual Review of Condensed Matter Physics, Vol 7* (eds. M. C. Marchetti & S. Sachdev), pp. 89-117.
[14] Wagner, A. 2007 From bit to it: The transformation of information into living matter by metabolic networks. . *BMC Systems Biology* **1**, 33.
[15] Bergstrom, C. T., Lachmann, M. & Ieee. 2004 *Shannon information and biological fitness*50-54 p.
[16] Frieden, B. R., Plastino, A. & Soffer, B. H. 2001 Population genetics from an information perspective. *Journal of Theoretical Biology* **208**, 49-64. (DOI:10.1006/jtbi.2000.2199).
[17] Weinberger, E. D. 2002 A theory of pragmatic information and its application to the quasi-species model of biological evolution. *Biosystems* **66**, 105-119. (DOI:10.1016/s0303-2647(02)00038-2).

[18] Iwasa, Y. 1988 Free fitness that always increases in evolution. *Journal of Theoretical Biology* **135**, 265-281.
[19] Day, T. 2015 Information entropy as a measure of genetic diversity and evolvability in colonization. *Molecular Ecology* **24**, 2073-2083.
[20] Wilson, D. S. & Szostak, J. W. 1999 In vitro selection of functional nucleic acids. *Annual Review of Biochemistry* **68**, 611-647.
[21] Keefe, A. D. & Szostak, J. W. 2001 Functional proteins from a random-sequence library. *Nature* **410**, 715-718.
[22] Curtis, E. A. & Bartel, D. P. 2013 Synthetic shuffling and in vitro selection reveal the rugged adaptive fitness landscape of a kinase ribozyme. *RNA* **19**, 1116-1128.
[23] Jiménez, J. I., Xulvi-Brunet, R., Campbell, G. W., Turk-MacLeod, R. & Chen, I. A. 2013 Comprehensive experimental fitness landscape and evolutionary network for small RNA. *Proceedings of the National Academy of Sciences of the U.S.A.* **110**, 14984-14989.
[24] Currin, A., Swainston, N., Day, P. J. & Kell, D. B. 2015 Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chemical Society Reviews* **44**, 1172-1239.
[25] Khersonsky, O., Rosenblat, M., Toker, L., Yacobson, S., Hugenmatter, A., Silman, I., Sussman, J. L., Aviram, M. & Tawfik, D. S. 2009 Directed evolution of serum paraoxonase PON3 by family shuffling and ancestor/consensus mutagenesis, and its biochemical characterization. *Biochemistry* **48**, 6644-6654.
[26] Romero, P. A. & Arnold, F. H. 2009 Exploring protein fitness landscapes by directed evolution. *Nature Reviews Molecular Cell Biology* **10**, 866-876. (DOI:10.1038/nrm2805).
[27] Hayden, E. J., Ferrada, E. & Wagner, A. 2011 Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. *Nature* **474**, 92-95. (DOI:10.1038/nature10083).
[28] Szendro, I. G., Schenk, M. F., Franke, J., Krug, J. & de Visser, J. 2013 Quantitative analyses of empirical fitness landscapes. *Journal of Statistical Mechanics-Theory and Experiment*, P01005. (DOI:10.1088/1742-5468/2013/01/p01005).
[29] Buckling, A., Maclean, R. C., Brockhurst, M. A. & Colegrave, N. 2009 The Beagle in a bottle. *Nature* **457**, 824-829.
[30] Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., Chan, E. T., Metzler, G., Vedenko, A., Chen, X. Y., et al. 2009 Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720-1723. (DOI:10.1126/science.1162327).
[31] Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I. & Cook, K. 2014 Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431-1443.
[32] Feist, A. M., Herrgard, M. J., Thiele, I., Reed, J. L. & Palsson, B. O. 2009 Reconstruction of biochemical networks in microorganisms. *Nature Reviews Microbiology* **7**, 129-143. (DOI:10.1038/nrmicro1949).
[33] Roscoe, B. P., Thayer, K. M., Zeldovich, K. B., Fushman, D. & Bolon, D. N. A. 2013 Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *Journal of Molecular Biology* **425**, 1363-1377. (DOI:10.1016/j.jmb.2013.01.032).
[34] Araya, C. L., Fowler, D. M., Chen, W. T., Muniez, I., Kelly, J. W. & Fields, S. 2012 A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 16858-16863. (DOI:10.1073/pnas.1209751109).

[35] Schuster, P., Fontana, W., Stadler, P. & Hofacker, I. 1994 From sequences to shapes and back - a case-study in RNA secondary structures. *Proceedings of the Royal Society of London Series B* **255**, 279-284.
[36] Lipman, D. & Wilbur, W. 1991 Modeling neutral and selective evolution of protein folding. *Proceedings of the Royal Society of London Series B* **245**, 7-11.
[37] Rodrigues, J. F. M. & Wagner, A. 2009 Evolutionary plasticity and innovations in complex metabolic reaction networks. *PLoS Computational Biology* **5**. (DOI:10.1371/journal.pcbi.1000613).
[38] Payne, J. L. & Wagner, A. 2014 The robustness and evolvability of transcription factor binding sites. *Science* **343**, 875-877.
[39] Hosseini, S.-R., Barve, A. & Wagner, A. 2015 Exhaustive analysis of a genotype space comprising 10^{15} central carbon metabolisms reveals an organization conducive to metabolic innovation. *PLoS Comput Biol* **11**, e1004329.
[40] Wagner, A. 2008 Robustness and evolvability: a paradox resolved. *Proceedings of the Royal Society of London Series B-Biological Sciences.* **275**, 91-100.
[41] Manrubia, S. & Cuesta, J. A. 2015 Evolution on neutral networks accelerates the ticking rate of the molecular clock. *Journal of the Royal Society Interface* **12**, 20141010.
[42] Ancel, L. W. & Fontana, W. 2000 Plasticity, evolvability, and modularity in RNA. *Journal of Experimental Zoology/Molecular Development and Evolution* **288**, 242-283.
[43] Ferrada, E. & Wagner, A. 2008 Protein robustness promotes evolutionary innovations on large evolutionary time scales. *Proceedings of the Royal Society of London Series B-Biological Sciences.* **275**, 1595-1602.
[44] Greenbury, S. F., Johnston, I. G., Louis, A. A. & Ahnert, S. E. 2014 A tractable genotype–phenotype map modelling the self-assembly of protein quaternary structure. *Journal of the Royal Society Interface* **11**, 20140249.
[45] Cover, T. M. & Thomas, J. A. 2006 *Elements of information theory*. 2nd ed. Wiley, Hoboken, New Jersey.
[46] Wootton, J. C. 1994 Non-globular domains in protein sequences: automated segmentation using complexity measures. *Computers & Chemistry* **18**, 269-285.
[47] Szostak, J. W. 2003 Molecular messages. *Nature* **423**, 689-689. (DOI:10.1038/423689a).
[48] Adami, C., Ofria, C. & Collirer, T. C. 2000 Evolution of biological complexity. *Proceedings of the National Academy of Sciences* **97**, 4463-4468.
[49] Ohno, S. 1970 *Evolution by gene duplication*. New York, Springer.
[50] Theissen, G. 2001 Development of floral organ identity: stories from the MADS house. *Current Opinion in Plant Biology* **4**, 75-85.
[51] Wagner, A. 2008 Gene duplications, robustness and evolutionary innovations. *Bioessays* **30**, 367-373.
[52] Wray, G. A. 2007 The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics* **8**, 206-216. (DOI:10.1038/nrg2063).
[53] Prud'homme, B., Gompel, N. & Carroll, S. B. 2007 Emerging principles of regulatory evolution. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 8605-8612. (DOI:10.1073/pnas.0700488104).
[54] Carroll, S. B., Grenier, J. K. & Weatherbee, S. D. 2001 *From DNA to diversity. Molecular genetics and the evolution of animal design*. Malden, MA, Blackwell.
[55] Tuğrul, M., Paixão, T., Barton, N. H. & Tkačik, G. 2015 Dynamics of transcription factor binding site evolution. *PLoS Genet* **11**, e1005639.

[56] Berg, J., Willmann, S. & Lässig, M. 2004 Adaptive evolution of transcription factor binding sites. *BMC Evolutionary Biology* **4**, 1.
[57] Weaver, D. C., Workman, C. T. & Stormo, G. D. 1999 Modeling regulatory networks with weight matrices. *Pacific Symposium on Biocomputing* **4**, 112-123.
[58] Carothers, J. M., Oestreich, S. C., Davis, J. H. & Szostak, J. W. 2004 Informational complexity and functional activity of RNA structures. *Journal of the American Chemical Society* **126**, 5130-5137. (DOI:10.1021/ja031504a).
[59] Becker, S. A., Feist, A. M., Mo, M. L., Hannum, G., Palsson, B. O. & Herrgard, M. J. 2007 Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nature Protocols* **2**, 727-738. (DOI:10.1038/nprot.2007.99).
[60] Edwards, J. S., Ibarra, R. U. & Palsson, B. O. 2001 In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nature Biotechnology* **19**, 125-130.
[61] Segre, D., Vitkup, D. & Church, G. 2002 Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences of the U.S.A.* **99**, 15112-15117.
[62] Forster, J., Famili, I., Fu, P., Palsson, B. & Nielsen, J. 2003 Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Research* **13**, 244-253.
[63] Wang, Z. & Zhang, J. Z. 2009 Abundant indispensable redundancies in cellular metabolic networks. *Genome Biology and Evolution* **1**, 23-33. (DOI:10.1093/gbe/evp002).
[64] Hosseini, S.-R., Barve, A. & Wagner, A. 2015 Exhaustive analysis of a genotype space comprising 10^{15} central carbon metabolisms reveals an organization conducive to metabolic innovation. *PLoS Computational Biology* **11**, e1004329.
[65] Hartl, D. L. & Clark, A. G. 2007 *Principles of population genetics*. 4th ed. Sunderland, MA, Sinauer Associates.
[66] Kimura, M. 1983 *The neutral theory of molecular evolution*. Cambridge, Cambridge University Press.
[67] Ewens, W. J. 2012 *Mathematical Population Genetics 1: Theoretical Introduction*. New York, NY, Springer Science & Business Media.
[68] Fontana, W. & Schuster, P. 1998 Continuity in evolution: On the nature of transitions. *Science* **280**, 1451-1455.
[69] Strelioff, C. C., Lenski, R. E. & Ofria, C. 2010 Evolutionary dynamics, epistatic interactions, and biological information. *Journal of Theoretical Biology* **266**, 584-594. (DOI:10.1016/j.jtbi.2010.07.025).
[70] Lee, H., Popodi, E., Tang, H. X. & Foster, P. L. 2012 Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **109**, E2774-E2783. (DOI:10.1073/pnas.1210309109).
[71] Schaper, S., Johnston, I. G. & Louis, A. A. 2012 Epistasis can lead to fragmented neutral spaces and contingency in evolution. *Proceedings of the Royal Society B-Biological Sciences* **279**, 1777-1783. (DOI:10.1098/rspb.2011.2183).
[72] Muller, G. B. & Wagner, G. P. 1991 Novelty in evolution: restructuring the concept. *Annual Review of Ecology and Systematics*, 229-256.
[73] Aguilar-Rodriguez, J. P., J.A. & Wagner, A. 2017 1000 empirical adaptive landscapes and their navigability. (in press). *Nature Ecology and Evolution*.